

Revealing effective classifiers through network comparison

Lazaros K. Gallos and Nina H. Fefferman

*Department of Ecology, Evolution, & Natural Resources,
Rutgers University, New Brunswick NJ 08901, USA,
and DIMACS, Rutgers University, Piscataway NJ 08854, USA*

(Dated: November 27, 2014)

Abstract

The ability to compare complex systems can provide new insight into the fundamental nature of the processes captured in ways that are otherwise inaccessible to observation. Here, we introduce the n -tangle method to directly compare two networks for structural similarity, based on the distribution of edge density in network subgraphs. We demonstrate that this method can efficiently introduce comparative analysis into network science and opens the road for many new applications. For example, we show how the construction of a phylogenetic tree across animal taxa according to their social structure can reveal commonalities in the behavioral ecology of the populations, or how students create similar networks according to the University size. Our method can be expanded to study a multitude of additional properties, such as network classification, changes during time evolution, convergence of growth models, and detection of structural changes during damage.

PACS numbers: 89.75.Fb, 89.75.Da, 87.23.Ge

Advances in quantitative methods for network analysis have allowed researchers across fields to quantify and characterize patterns of interaction among individuals, with applications in a startling diversity of fields[1]. As in the progression of many quantitative tools, while initial efforts to use network analysis were mainly descriptive[2], research then advanced to focus on using them as predictive tools, isolating particular characteristics that can provide insight into the system of interest[3, 4]. However, the richest and most interesting level of investigation from new metrics frequently arises when they are ultimately used to make comparisons across systems, discovering which characteristics are shared and which are not. The ability to compare systems has always been a strong driving force in science[5]. Even the most straightforward new tests, such as the discovery of Gram staining for the classification of bacterial walls, can lead to breakthroughs that influence generations of research (in this case, becoming the cornerstone for progress in drug discovery and antibiotic therapies[6]).

Currently, there is not a rigorous definition of network similarity. This allows similarity to be as broadly interpreted as just one single quantity averaged over the entire system - e.g. networks with the same average degree - or it can be extremely restrictive, e.g. node-to-node correspondence in identical networks. Obviously, no one property can fully characterize a network: for instance, networks can be structurally very different if they have the same degree distribution but different clustering coefficient. Even if the clustering coefficient is the same, it is possible that the networks will have different modularity, etc. It is not known how many and which properties should be combined to construct a weighted index of similarity. Therefore, current research has been directed to alternative methods. Motif comparison [7] or graphlet comparison [8], for example, is based on the idea that if we continuously isolate parts of the network and find the same patterns to occur in the same frequency in two networks, these networks will have a higher probability of being ‘similar’ to each other. However, there are many practical constraints that render these techniques incapable of handling larger networks or larger motifs [9]. The most recent advance in the field [10] introduced a novel concept in which the network is broken down in communities at different scales and the comparison is based on network modularity properties. The question of similarity under this method becomes ‘how similar the modular structure of the networks is’.

Here we introduce a measure to detect similarity based on direct topological properties:

Topological Analysis of Network subGraph Link/Edge (tangle) Density. Many of these properties can be captured by the distance from a tree structure at different length scales. The method combines the insight of motifs, simplified for efficiency, and focuses on microscopic structure compared to the mesoscopic approach of modularity comparison in Onnela et al [10]. Where the advantage of the motif method is that it takes into account the local configuration of the links, if we relax the motif requirement for exactly matching patterns we can use the links density as our metric. The basic foundation of our method is to calculate how the density of links behaves at different scales across the network. This choice represents many advantages since it incorporates information from many structural properties, e.g. on the degree (naturally through $\rho = \langle k \rangle / n$), the clustering coefficient (n -tangle=3 in our method), the number of loops (every additional link in a tree structure increases the number of loops), etc. Additionally the calculation of the similarity index is straightforward and is bounded between 0 and 1.

It is possible to use other properties instead of density, such as the local degree distribution or clustering coefficient, but the crucial step is the sampling of the connected subgraphs. For example, a specific partition of the network, such as one optimizing modularity [11] does not contain enough information about the network structure. Similarly, network-wide metrics cannot capture the local details, the natural inhomogeneity, or possible scale-dependent differences in structure.

Notice that our definition of similarity is ‘similar local structures’, or equivalently that the extended neighborhood of a node looks similar with the extended neighborhood of another node. In this approach, similarity indicates the number of loops that exist in continuously expanding scales. Therefore, all tree structures will be deemed equivalent by our method, even if they are different structures, e.g. a scale-free tree vs an ER tree. In other words, our question for similarity becomes ‘how far away is a given structure from a tree’ or equivalently ‘how close is it to a complete subgraph’. This question is easily calculable and takes into account possible local deviations of the local structure from the global topology.

The crux of this method is to capture how many affiliations we expect to find when we isolate any given size of connected sub-group. The concept is the following: Consider a connected group of 10 students, which is randomly selected from a class of 100 students. If you are in this group, how many direct friends do you expect to find in this sample, or in other words what is the average edge density in the group? We define this to be the

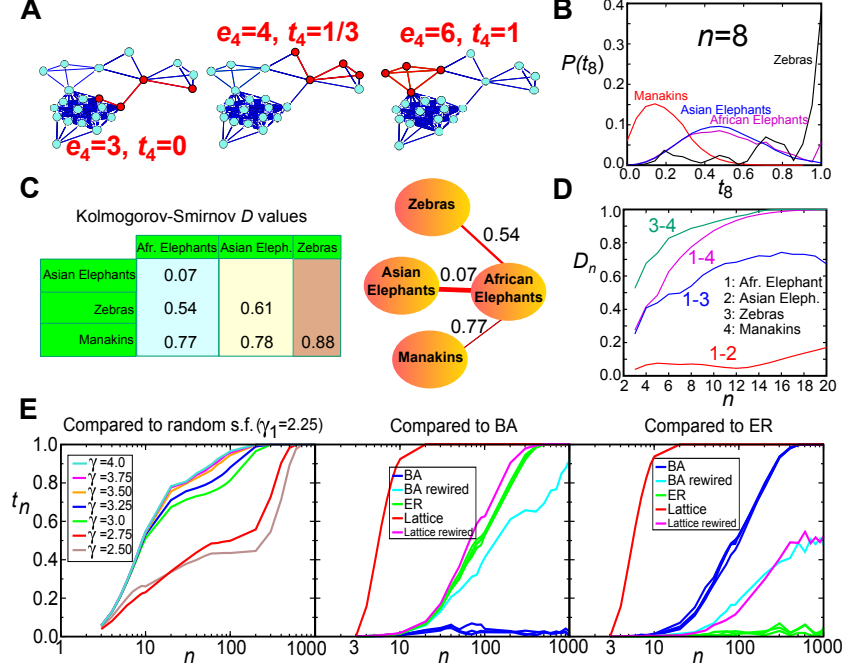


FIG. 1: The n -tangle method. (A) We randomly sample connected induced subgraphs of n nodes and calculate their normalized link density t_n . (B) We construct the n -tangle histogram $P(t_n)$ for a given value of n (the example shows the 8-tangle distribution for 4 animal social networks). (C) We calculate the distance between any two distributions I and J through, e.g., a Kolmogorov-Smirnov statistic $D_n(I - J)$. These D values are used as the distance between the original networks and can be mapped to a minimum spanning tree (shown here for the four networks), a hierarchical tree, or a threshold-based network. (D) Variation of the distance between these 4 networks as a function of the subgraph size n . (E) The distance of a random scale-free network with a degree exponent γ from a network with $\gamma_1 = 2.25$ increases monotonically as we increase the value of γ (left). Similarly, the n -tangle distance among random Barabasi-Albert networks (center) or random Erdos-Renyi networks (right) is close to zero, while distances with other model networks are significantly higher.

10-tangle density (or n -tangle, for any n). If we construct the histogram of densities from different samples then we can compare these distributions in two different networks, and we can know the extent of association in a group of a given size independently of the pattern formed in each subgroup. In this way, our method bypasses the need to determine direct node-to-node correspondence[12], while still capturing node-level properties of the network for comparison.

Formally, we define the n -tangle method in the following way. In a graph $G(V, E)$ comprising a set V of nodes and a set E of edges we isolate all possible connected induced sub-graphs $G^{\text{in}}(V_n, E_n)$. The condition for these sub-graphs is that they should include exactly n nodes ($|V_n| = n$) and the subset E_n of E should include all e_n links among those n nodes in G . For each subgraph we define the n -tangle density, t_n , as the normalized edge density of this subgraph, i.e. the fraction of existing over all possible links, after we remove the $n - 1$ links that are needed for connectivity:

$$t_n = \frac{e_n - (n - 1)}{n(n - 1)/2 - (n - 1)} = \frac{2(e_n - n + 1)}{n^2 - 3n + 2}. \quad (1)$$

It is important that the size of the n -tangle remains much smaller than the network size N , $n \ll N$, so that the sampled subgraphs are statistically independent from each other. To include the considerably inhomogeneous character of the local structure in networks, we consider the n -tangle distribution $P(t_n)$ of all G^{in} (Figs 1A, 1B). This distribution represents the ‘signature’ of a network at a given subgraph size n . We repeat this process for all different subgraph sizes n , resulting to potentially different signatures as we vary n . We can then compare the degree of similarity of two networks A and B at a given scale by a simple Kolmogorov-Smirnov statistic[13], $D_n(A - B) = \sup |F_A(t_n) - F_B(t_n)|$, where $F_A(t_n)$ is the corresponding n -tangle cumulative probability in network A and \sup denotes the supremum value (Fig. 1C). Other metrics of distribution distance can also be used with similar results, (see Supplemental Material). Since the full comparison involves all subgraph sizes, this method can reveal how two networks can be similar at a local scale, while at a larger scale they may exhibit different structures, allowing both global network comparison and local analysis of the scale at which similarity may be greatest (Fig. 1D).

Our approach avoids the inherent constraints of motif[7] or graphlet[8] based methods[14], by ignoring the costly calculation of the specific pattern created by the group and instead placing emphasis on the density of the group, i.e. a single number. Therefore, the exponential increase in the number of patterns as a function of group size, which limits those techniques to very small-size patterns, does not influence the applicability of our method to larger sub-graphs. Now, we only need to keep the number of links for each configuration, which makes the calculation and storage very fast. Even though the computational complexity of the n -tangle method does increase with the subgraph size, the connectedness of bigger social groups can be probed at practically any size n , through a fast sampling method. We

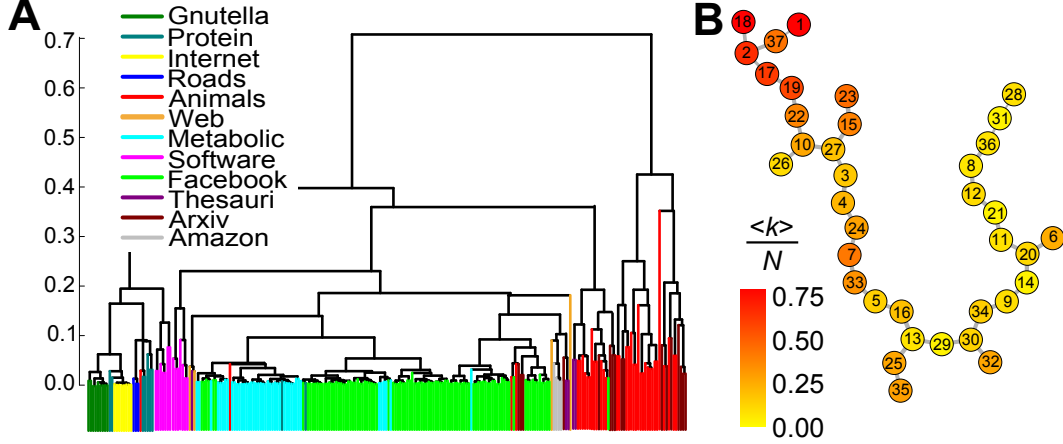


FIG. 2: (A) Hierarchical tree of 236 networks from different fields, based on the n -tangle distance (here $n=5$). We used the UPGMA (Unweighted Pair-Group Method using Arithmetic Averages) hierarchical clustering method[25]. Colors represent networks in the same family, as indicated in the index. (B) The Minimum Spanning Tree for animal networks, based on the n -tangle distance ($n = 12$). The species color corresponds to varying levels of normalized degree $\langle k \rangle / N$ and separates nicely the species. The node numbers correspond to the species in table I of the SM.

used a simple Monte-Carlo method to sample a large number of configurations, where we repeatedly selected random subgraphs and calculated the links within the subgraph.

Our method is designed to quantify local edge densities which combine a lot of structural information, and as such it can successfully detect changes in standard network properties. In Fig. 1E we compare a series of random scale-free networks created by the configuration model with a similar network with degree exponent $\gamma_1 = 2.25$. The networks become more distant as the exponent of these networks increases, demonstrating that the method can separate similar structures with different parameters. Similarly, we compare a number of networks to a sample Barabasi-Albert (BA) network. Different realizations of BA networks are found to be at almost zero distance with each other, but a randomly rewired BA network has a different structure. Similarly, lattices and ER networks are also far from the BA network. Analogous results are found when we compare these model networks with an ER network.

We demonstrate the n -tangle method first by comparing 236 network structures of different origins (described in the Supplemental Material). The hierarchical tree in Fig. 2A indicates that networks from the same family tend to cluster with each other. For example,

friendship networks in facebook are in general closer to each other than with e.g. animal social networks, which also tend to be detected as similar. We consider this natural separation as a simple verification test for the method.

A more interesting problem is to detect network similarities in systems from within the same family. For example, we can construct a phylogeny of animal species based on their social structure[15]. In this way, we explore whether species with similar descriptive characterizations in behavioral ecology do, in fact, exhibit similar social structures[16]. We analyze empirically determined contact affiliation networks of 33 animal species (described in the Supplemental Material). In molecular biology, phylogenetic trees can be constructed from evolutionary distance[17] (pair wise distances between sequences). Here, rather than using species genetic data, our input data are the pair-wise distances of the n -tangle method.

We are therefore able, using our analysis, to determine whether or not a meaningful cluster results from a choice of a particular facet of the system. In this example, we find that the normalized average degree, i.e. $\langle k \rangle / N$, is able to generate clear clustering by n -tangle analysis. This result of our method can provide the first insights into whether qualitatively similar social classifications in fact yield similar population-level networks of interaction across species (for example, do all dominance hierarchies yield similar social structures for the entire population?). This is a critical next step in understanding animal social systems.

The n -tangle method can also be used to isolate key network features that enable classification of networks. In Fig. 3 we present the n -tangle connectivity trees resulting from a) Facebook friendship networks in 100 Universities in 2005[18] (described in the Supplemental Materials), b) arxiv.org co-authorship[19] in 17 different fields, and c) software code in 14 different projects[20, 21]. For the Facebook friendship, there is no clear clustering with the average degree, but when we consider student enrollment, then we discover a similarity between networks at Universities of similar size, at all sizes. The n -tangle method therefore enables us to obtain meaningful sociological insight into the process, where students create online friendships according to the size of the pool of possible connections, even though the average number of friendships is much smaller than the pool size. It may therefore be that the fundamental nature of the social activities and experience is shaped by the total size of the university, even though that number can be significantly larger than the size of the average friend-group. For the case of co-authorship, on the other hand, the classification of networks according to the network size does not work well. We instead discover that the

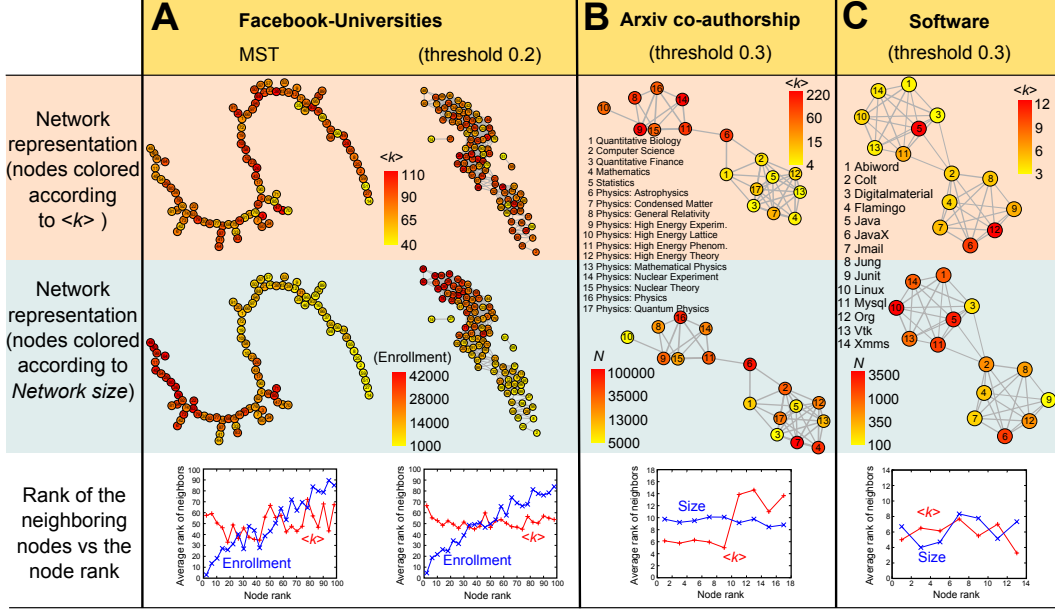


FIG. 3: Comparison of static networks. (A) Minimum Spanning Tree and threshold-based network representation of similarities in the networks of facebook friendship in 100 US Universities. The color of the University-nodes corresponds to either the average degree or the University size, in terms of enrollment size. The enrollment size is the key property for clustering. The plot at the bottom row compares the rank of a University to its neighbors rank. The enrollment size has a very hierarchical structure where ranks of the same order connect to each other, in contrast to average degree ranking where a nodes rank cannot predict the rank of its neighbors. (B) The similarity network of scientific fields, based on co-authorship, exhibits the opposite trend. The average degree is a nice indicator for clustering, while the network size is not. This result is supported by the plot comparing the rank with the neighbors rank. (C) The network of similarity between software projects cannot be clustered according to either the average degree or the network size. The two modules correspond however to networks that were built by two independent methods.

important factor in this case is the average degree of an author, i.e. fields with large number of co-authors yield similar networks with each other. This classification of networks according to an underlying structural property does not trivially result from the n -tangle method. In the example of software project networks in Fig. 3c we were not able to determine any particular structural property that separates the projects in the n -tangle networks. Interestingly, each of the two modules in Fig. 3c includes software projects that were generated by different circumstances. This method therefore, not only allows comparison across networks,

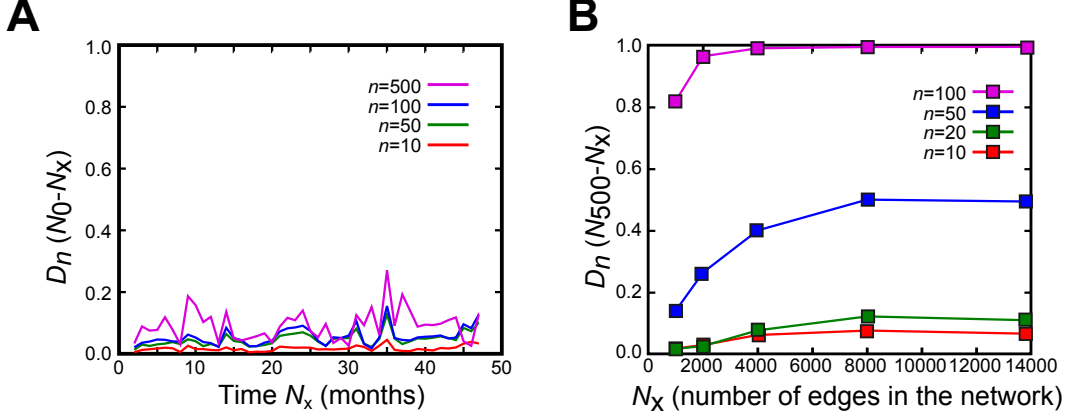


FIG. 4: Comparison of network evolution. (A) Similarity of the Internet at the AS level with itself as a function of time. We compare the KS index $D_n(N_x - N_0)$ of the Internet topology at N_0 = January 2004 with the topology at time N_x , which is increasing monthly. Independently of the scale n , the topology remains the same throughout the network evolution for 3 years. (B) Comparison of the KS index in social networking friendships as a function of time. We compare the network topology of the early network containing 500 links with the networks at subsequent times. The network remains the same for small values of n , but changes drastically at larger scales.

but enables hypothesis testing about which facets might be the most salient organizational features that drive the emergence of networks within the systems studied.

We also applied this method to characterize network evolution. In the examples of the Internet growth[22] and online social-networking evolution[23] in Fig. 4, we compare the network at a given time with the same network at subsequent times. The starting date for the Internet data was January 2004. Our method indicates that the Internet topology was already fixed in time by January 2004 and did not change much by November 2007, when the network had already doubled in size. This result holds across all subgraph sizes, and is also consistent with the macroscopic fact that the average link density was declining slowly from $2.4 \cdot 10^{-4}$ to $1.4 \cdot 10^{-4}$ over three years. On the contrary, the facebook-like online network shows a stable behavior only at small scales n . The number of edges in the network increases by a factor of 25, but the n -tangle density remains very similar at any time when $n < 20$. When we consider larger n values, though, there is a very sharp change between the initial reference network and the subsequent instances of it. Therefore, within the same network the small-scale structures remain the same, while larger-scale structures evolve into

different forms. The method can therefore separate structurally stabilized networks over time from unstabilized ones. Moreover, in networks of evolving topology we can identify differences in the stability of short-scale and larger-scale structures. This may therefore enable accurate estimation of the quality of approximation of static snapshots of continually shifting networks, which has been shown to be of critical importance in areas such as epidemiology[24].

The calculation of the n -tangle density provides a simple and powerful method for efficient network comparison. Understanding the degree of similarity between two networks is the key to promote the classification of networks into clusters for further analysis of their common features that would otherwise remain unknown, and allows us to hypothesize meaningfully about how these clusters may capture fundamental properties of networks and the systems they represent.

We thank the Dept. of Homeland Security for funds in support of this research through the CCICADA Center at Rutgers, and NSF EaeM grant #1049088. We are grateful to the following people, for making their animal network data available to us: C. Berman, J.R. Madden, J. Wolf, D.B. McDonald, P.C. Cross, and S. Sundaresan.

-
- [1] A.-L. Barabasi, *Nature Physics* **8** 14 (2011).
 - [2] R. Albert and A.-L. Barabasi, *Rev. Mod. Phys.* **74** 47 (2002).
 - [3] A. Clauset, C. Moore, and M. E. J. Newman, *Nature* **453** 98 (2008).
 - [4] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. A. Makse, *Nature Physics* **6** 888 (2010).
 - [5] C. R. Woese, O. Kandler, and M. L. Wheelis, *Proc. Natl. Acad. Sci. USA* **87** 4576 (1990).
 - [6] A. Popescu and R. J. Doyle, *Biotechnic & histochemistry* **71** 145 (1996).
 - [7] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, *Science* **298** 824 (2002).
 - [8] N. Przulj, *Bioinformatics* **23** e177 (2007).
 - [9] K. Baskerville and M. Paczuski, *Phys. Rev. E* **74** 051903 (2006).
 - [10] J.-P. Onnela, D.J. Fenn, S. Reid, M.A. Porter, P.J. Mucha, M.D. Fricker, and N.S. Jones, *Phys. Rev. E* **86** 036104 (2012).

- [11] Y.-Y. Ahn, J.P. Bagrow, and S. Lehman, *Nature* **466** 761 (2010).
- [12] P. Papadimitriou, A. Dasdan, and H. Garcia-Molina, *Journal of Internet Services and Applications* **1**, 19 (2010).
- [13] M. A. Stephens, *Journal of the American Statistical Association* **69** 730 (1974).
- [14] J. Ugander, L. Backstrom, and J. Kleinberg, *Proc. 22nd International World Wide Web Conference* (2013).
- [15] R. D. Alexander, *Annual Review of Ecology and Systematics* **5** 325 (1974).
- [16] D. B. McDonald, *Proc. Natl. Acad. Sci. USA* **104** 10910 (2007).
- [17] B. PrudHomme, N. Gompel, A. Rokas, V. A. Kassner, T. M. Williams, S.-D. Yeh, J. R. True, and S. B. Carroll, *Nature* **440** 1050 (2006).
- [18] A. L. Traud, P. J. Mucha, M. A. Porter, *Physica A* **391** 4165 (2012).
- [19] M. E. J. Newman, *Proc. Natl. Acad. Sci. USA* **98** 404 (2001).
- [20] C. R. Myers, *Phys. Rev. E* **68** 046116 (2003).
- [21] L. Subelj and M. Bajec, *Physica A* **390** 2968 (2011).
- [22] J. Leskovec, J. Kleinberg, and C. Faloutsos, *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 177 (2005).
- [23] T. Opsahl and P. Panzarasa, *Social networks* **31** 155 (2009).
- [24] N. H. Fefferman and K. L. Ng, *Phys. Rev. E* **76** 031919 (2007).
- [25] C. D. Michener and R. R. Sokal, *Evolution* 130 (1957).

SUPPLEMENTARY INFORMATION

Revealing effective classifiers through network comparison

Lazaros K. Gallos and Nina H. Fefferman

DATASETS

In our study we have used a number of datasets to build and analyze networks. A short description of these datasets and their sources is as follows:

a) *Animal affiliation networks*. We have compiled a set of 37 empirically determined ‘social’ networks in different species that can be found in the published literature. Association among individuals was defined by affiliative behavior, such as proximity, grooming interactions, etc. Our analysis included primate populations, herd mammals, marine mammals, fish, birds, insects, and reptiles. A detailed list of the species used, along with the references from where the networks were extracted, is shown in Table 1. The network sizes varied from 18 to 380 individuals. In some cases, we had more than one network for each species, depending on when the network was recorded. These cases are indicated in the nodes column of Table 1.

b) *Facebook in 100 Universities*. These networks are based on facebook friendship connections in 100 Colleges and Universities in USA on September 2005. The data have been made publically available [32] and have been analyzed in Ref. [33]. At the early stages of facebook, only students from specific Universities could create accounts in the site. The data that we use were recorded for the 100 first Universities that joined facebook, and contain friendships only within the same University, providing us with 100 independent networks. The networks are complete, in the sense that they contain all the existing nodes and all the connections at that time.

We isolated the largest clusters in each case. The University IDs that are used in Fig. 3a of the main text are shown in Table 2. In the same table we provide the values of the average degree and the student enrollment size of each University in 2005, which are used for the color coding of Fig. 3a. We downloaded the enrollment information from the datacenter of the *National Center for Education Statistics*: <http://nces.ed.gov/ipeds/datacenter/> by creating a custom query for the 12-month full-time equivalent enrollment in the Academic year 2005-06.

c) *Arxiv co-authorship networks*. We downloaded the entire database of all papers submitted to arxiv.org from the beginning of the site in 1991 until December 31, 2012, using the Open Archive Initiative (oai2) protocol. The databases were parsed to identify unique authors and the authors of each paper. We used the site's classification of papers into 18 broad categories, and created one network for each category using all the papers in that field. The network nodes correspond to authors and a link suggests that these two nodes have co-authored at least one paper in this category. The 18 fields were the following, with the size of the largest cluster in the network shown in parentheses: Quantitative Biology (8618), Computer Science (30689), Quantitative Finance (1514), Math (52351), Statistics (6360), Physics / Astrophysics (71155), Physics / Condensed Matter (85956), Physics / General Relativity (16189), Physics / High Energy Physics -Experiment (28806), Physics / High Energy Physics - Lattice (5130), Physics / High Energy Physics - Phenomenology (31991), Physics / High Energy Physics - Theory (21398), Physics / Mathematical Physics (10918), Physics / Nuclear Experiments (19933), Physics / Nuclear Theory (14121), Physics / Physics (61665), Physics / Quantum Physics (22926), Physics / Nonlinear Sciences (13). Due to the small size of the largest cluster in the Nonlinear Sciences field we did not consider this network in our study.

d) *Software networks*: We used networks of software from two sources: 1) The data for *junit*, *jmail*, *flamingo*, *jung*, *colt*, *org*, *java*, and *javax* were downloaded from <http://lovro.lpt.fri.uni-lj.si/publications.jsp?show=ssc> . The analysis of these data was done in [34]. The network is created by connections between the classes - nodes - in each software code. Two classes are considered to be connected through the following dependencies: inheritance, field, parameter and return. 2) We also used the software packages *Abiword*, *DigitalMaterial*, *Linux*, *Mysql*, *VTK*, and *XMMS* from Ref. [35], that were also constructed according to class collaboration.

e) *Internet* (evolving network). We downloaded the CAIDA Autonomous System graphs from January 2004 to November 2007 from the SNAP Stanford datasets in <http://snap.stanford.edu/data/as-caida.html> . The data are described and analyzed in Ref. [36]. Starting from January 2004, we used a total of 47 static snapshots which were roughly one month apart each. This allows us to monitor the evolution of the network over 3 years.

f) *Messages in an online social networking site* (evolving network). This dataset was downloaded from http://toreopsahl.com/datasets/#online_social_network and has been analyzed in [37]. It corresponds to online messages sent among students at the University of California, Irvine, through a “Facebook-like Social Network”. The original form of the network was directed, so we projected it to an undirected form by ignoring the directionality of the links. Each message was time-stamped so we were able to follow the entire network evolution. Our starting point was when the first 500 links were created, and we sampled snapshots of the network with 1000, 2000, 4000, 8000, and 13838 links.

g) *Gnutella sharing*. We used the 9 snapshots of a peer-to-peer Gnutella network [38], where nodes represent hosts and links are the connections between these hosts. This is a directed network, so we used its undirected projection.

h) *Protein Interaction networks*. We used 9 protein interaction networks from BioGrid [39] for the following species: *A. thaliana*, *C. elegans*, *D. melanogaster*, *H. sapiens*, *M. musculus*, *P. falciparum*, *R. norvegicus*, *S. cerevisiae*, and *S. pombe*.

i) *Metabolic networks*. We used the 43 metabolic networks from Ref. [40].

j) *Road networks*. A node in this network represents an intersection and the links correspond to the roads that connect these intersections. The three state-wide networks we used were for California, Pennsylvania, and Texas [41].

k) *Thesauri networks*. We extracted the networks from 5 thesauri datasets, where nodes represent words and the links indicate that the two words are synonyms. These data were extracted from the LibreOffice Thesaurus and correspond to the following languages: English (UK), English (US), Spanish (AR), Spanish (ES), and Spanish (VE).

l) *Web networks*. The nodes in these networks represent webpages and the network links represent hyperlinks connecting these webpages [41]. We converted all links to undirected, and used the web network of Berkeley and Stanford, Google, Notre Dame, and Stanford. The datasets were downloaded from the Stanford SNAP database.

m) *Amazon co-purchase*. This network connects items that were frequently purchased together in amazon.com, as found by crawling software [42]. The networks were converted to undirected. We used four networks based on data collected on 2003 on these dates: March 2, March 12, May 5, June 1. The data were downloaded from the Stanford SNAP database.

TABLE S1

COMMON NAME	SCIENTIFIC NAME	POPULATION	CITATION	NOTES
1. African Buffalo A	<i>Syncerus Cafer</i>	39	[1]	Network in May 2002
2. African Buffalo B	<i>Syncerus Cafer</i>	64	[1]	Aggregate Nov 01-Oct 03
3. African Elephant	<i>Loxodonta Africana</i>	112	[2]	
4. Asian Elephant	<i>Elephas Maximus</i>	105	[2]	
5. Bats	<i>Thyroptera Tricolor</i>	55	[3]	
6. Brushtail Possums	<i>Trichosurus vulpecula</i>	18	[4]	
7. Chimpanzees	<i>Pan troglodytes schweinfurthii</i>	18	[5]	
8. Cichlids	<i>Neolamprologous pulcher</i>	72	[6]	
9. Columbian Squirrel	<i>Spermophilus Columbianus</i>	65	[7]	
10. Crows	<i>Corvus moneduloides</i>	34	[8]	
11. Dolphins	<i>Tursiops truncatus</i>	62	[9]	
12. Fungus Beetle	<i>Bolitotherus cornutus</i>	34	[10]	
13. Giraffes	<i>Giraffa camelopardalis reticulata</i>	77	[11,23]	
14. Great Tits	<i>Parus Major</i>	104	[12]	
15. Guiana Dolphins	<i>Sotalia Guianensis</i>	49	[13]	
16. Guppies	<i>Poecilia reticulata</i>	63	[14]	
17. Hyenas A	<i>Crocota Crocuta</i>	35	[15]	Low-prey period 1999
18. Hyenas B	<i>Crocota Crocuta</i>	35	[15]	High-prey period 1999
19. Hyenas C	<i>Crocota Crocuta</i>	35	[15]	Low-prey period 2000
20. Lizards	<i>Egernia stokesii</i>	37	[16]	
21. Longtailed Manakins	<i>Chiroxiphia linearis</i>	156	[17]	
22. Marmot Meadows	<i>Marmota flaviventris</i>	22	[18]	
23. Onagers	<i>Equus hemionus khur</i>	28	[19]	
24. Orca	<i>Orcinus Orca</i>	43	[20]	
25. Pigtailed Macaques	<i>Macaca nemestrina</i>	48	[21]	
26. Pygmy Whales	<i>Feresa Attenuata</i>	103	[22]	
27. Red Deer	<i>Cervus Elaphus</i>	45	[23]	
28. Rhesus Macaques	<i>Macaca mulatta</i>	23	[24]	
29. Sea Lions	<i>Zalophus wollebaeki</i>	380	[25]	
30. Snubnosed Monkeys	<i>Rhinopithecus roxellana</i>	58	[26]	
31. Sparrows	<i>Melospiza melodia</i>	74	[27]	
32. Tasmanian Devil A	<i>Sarcophilus harrisii</i>	27	[28]	During mating season
33. Tasmanian Devil B	<i>Sarcophilus harrisii</i>	27	[28]	After mating season
34. Wild Baboons	<i>Papio anubis – Papio hamadryas</i>	35	[29]	
35. Wild Meerkats	<i>Suricatta suricatta</i>	24	[30]	
36. Wiretailed Manakins	<i>Pipra Filicauda</i>	46	[31]	
37. Zebras	<i>Equus grevyi</i>	23	[19]	

TABLE S2

ID	UNIVERSITY	ENROLLMENT	<k>	ID	UNIVERSITY	ENROLLMENT	<k>
1	American	10007	68.3	51	Reed	1201	39.1
2	Amherst	1642	81.4	52	Rice	5321	90.5
3	Auburn	23537	105.6	53	Rochester	11364	70.8
4	Baylor	14771	106.2	54	Rutgers	31563	63.9
5	Boston College	14328	84.7	55	Santa Clara	8075	84.8
6	Berkeley	33901	74.4	56	Simmons	4184	43.7
7	Bingham	13588	72.6	57	Smith	3118	65.4
8	Bowdoin	1666	75.0	58	Stanford	14087	98.1
9	Brandeis	5826	70.8	59	Swarthmore	1452	73.7
10	Brown	7904	89.6	60	Syracuse	16955	79.8
11	Boston University	30202	64.8	61	Temple	32740	52.8
12	Bucknell	3645	83.1	62	Tennessee	17045	90.8
13	CalPoly San Luis Obispo	17618	62.5	63	U. Texas Austin	46135	77.3
14	Caltech	2178	43.7	64	Texas A&M	42566	87.5
15	Carnegie	9091	75.5	65	Trinity	2116	85.7
16	Colgate	2744	89.1	66	Tufts	11100	74.9
17	Columbia	21522	75.9	67	Tulane	8191	73.4
18	Cornell	19602	84.9	68	U. Calif. Davis	29394	62.2
19	Dartmouth	6071	79.2	69	U. Calif. Irvine	25781	64.4
20	Duke	16151	102.5	70	U. Calif. Riverside	16443	45.6
21	Emory	14149	88.6	71	UCF	39139	57.4
22	FSU	36615	74.6	72	UChicago	9840	63.4
23	Georgetown	16369	90.7	73	UCLA	36864	73.1
24	GWU	19630	77.2	74	UConn	20672	70.3
25	Hamilton	1805	83.4	75	U. Calif. Santa Barbara	21938	64.7
26	Harvard	26324	109.3	76	U. Calif. Santa Cruz	15265	50.0
27	Haverford	1111	82.4	77	U. Calif. San Diego	26291	59.3
28	Howard	11276	101.2	78	Univ. Florida	52396	83.5
29	Indiana	36366	87.8	79	UGA	32539	96.3
30	JMU	16737	69.0	80	U. Illinois	43572	82.1
31	Johns Hopkins	17154	72.4	81	U. Mass. Amherst	22601	62.9
32	Lehigh	6102	78.2	82	UNC Chapel Hill	25757	84.5
33	Maine	9241	53.7	83	UPenn	26686	92.2
34	Maryland	31280	71.5	84	USC Columbia	24467	92.0
35	Michigan Tech	6125	43.7	85	USF	37461	48.1
36	Michigan	39240	78.2	86	U. San Francisco	8599	48.8
37	Middlebury	2902	81.2	87	U. Virginia	22977	91.9
38	Mississippi	14681	116.2	88	Vanderbilt	10260	106.1
39	MIT	10079	78.5	89	Vassar	2451	77.7
40	MSU	45166	69.1	90	Vermont	10426	52.2
41	Marquette	10344	84.2	91	Villanova	9559	81.2
42	Northeastern	19999	55.1	92	Virginia	27840	65.5
43	Northwestern	16976	92.7	93	Wake	6662	104.1
44	NotreDame	10832	89.1	94	Washington U. St Louis	12197	95.1
45	NYU	39783	66.2	95	Wellesley	2505	63.9
46	Oberlin	2840	61.6	96	Wesleyan	3515	76.9
47	Oklahoma	23350	102.5	97	William & Mary	7638	82.3
48	Penn State	41602	65.6	98	Williams College	2076	81.1
49	Pepperdine	7403	88.4	99	Wisconsin	35337	70.2
50	Princeton	7095	89.2	100	Yale	11288	94.7

REFERENCES FOR THE SUPPLEMENTARY INFORMATION SECTION

- [1] P.C. Cross, J.O. Lloyd-Smith, J.A. Bowers, C.T. Hay, M. Hofmeyr, & W.M. Getz, *Annales Zoologici Fennici* 41, 879 (2004).
- [2] S. de Silva & G. Wittemyer, *Int. J. Primatol.* 33, 1125 (2012).
- [3] G. Chaverri, *Behav. Ecol. Sociobiol.* 64, 1619 (2010).
- [4] L.A.L. Corner, D.U. Pfeiffer, & R.S. Morris, *Preventive Veterinary Medicine* 59, 147 (2003).
- [5] C.M. Murray, *Int. J. Primatol.* 28, 853 (2007).
- [6] R. Schürch, S. Rothenberger, & D. Heg, *Phil. Trans. R. Soc. B* 365, 4089 (2010).
- [7] T.G. Manno, *Animal Behaviour* 75, 1221 (2008).
- [8] C. Rutz, Z.T. Burns, R. James, S.M.H. Ismar, J. Burt, B. Otis, J. Bowen, & J.J.H. St Clair, *Current Biology* 22, R669 (2012).
- [9] D. Lusseau, K. Schneider, O.J. Boisseau, P. Haase, E. Slooten, S.M. Dawson, *Behav. Ecol. Sociobiol.* 54, 396 (2003).
- [10] V.A. Formica, C.W. Wood, W.B. Larsen, R.E. Butterfield, M.E. Augat, H.Y. Hougen, & E.D. Brodie III, *J. Evol. Biol.* 25, 130 (2012).
- [11] B. Shorrocks & D.P. Croft, *Afr. J. Ecol.* 47, 374 (2009).
- [12] I. Psorakis, S.J. Roberts, I. Rezek, & B.C. Sheldon, *J. R. Soc. Interface* 9, 3055 (2012).
- [13] M. Cantor, L.L. Wedekin, P.R. Guimarães, F.G. Daura-Jorge, M.R. Rossi-Santos, & P.C. Simões-Lopes, *Animal Behaviour* 84, 641 (2012).
- [14] D.P. Croft, J. Krause, S.K. Darden, I.W. Ramnarine, J.J. Faria, & R. James, *Behav. Ecol. Sociobiol.* 63, 1495 (2009).
- [15] K.E. Holekamp, J.E. Smith, C.C. Strelhoff, R.C. Van Horn, & H.E. Watts, *Molecular Ecology* 21, 613 (2012).
- [16] S.S. Godfrey, C.M. Bull, R. James, & K. Murray, *Behav. Ecol. Sociobiol.* 63, 1045 (2009).
- [17] D.B. McDonald, *Proc. Natl. Acad. Sci USA* 104, 10910 (2007).
- [18] D.T. Blumstein, T.W. Wey, & K. Tang, *Proc. R. Soc. B* 276, 3007 (2009).
- [19] S.R. Sundaresan, I.R. Fischhoff, J. Dushoff, & D.I. Rubenstein, *Oecologia* 151, 140 (2007).
- [20] P.R. Guimarães, Jr., M.A. de Menezes, R.W. Baird, D. Lusseau, P. Guimarães, & S.F. dos Reis, *Phys. Rev. E* 76, 042901 (2007).

- [21] J.C. Flack, M. Girvan, F.B.M. de Waal, & D.C. Krakauer, *Nature* 439, 426 (2006).
- [22] D.J. McSweeney, R.W. Baird, S.D. Mahaffy, D.L. Webster, & G.S. Schorr, *Marine Mammal Science* 25, 557 (2009).
- [23] D.P. Croft, R. James, & J. Krause, *Exploring Animal Social Networks* (Princeton Univ. Press, 2008).
- [24] C.M. Berman, K.L.R. Rasmussen, & S.J. Suomi, *Animal Behaviour*, 53, 405 (1997).
- [25] J.B.W. Wolf & F. Trillmich, *Proc. Biol. Sci.* 275, 2063 (2008).
- [26] P. Zhang, B. Li, X. Qi, A.J.J. MacIntosh, & K. Watanabe, *Int. J. Primatol.* 33, 1081 (2012).
- [27] C.N. Templeton, V.A. Reed, S.E. Campbell, & M.D. Beecher, *Behavioral Ecology* 23, 141 (2012).
- [28] R.K. Hamede, J. Bashford, H. McCallum, & M. Jones, *Ecology Letters* 12, 1147 (2009).
- [29] K. Sugawara, *Primates* 20, 21 (1979).
- [30] J.R. Madden, J.A. Drewe, G.P. Pearce, & T.H. Clutton-Brock, *Behav. Ecol. Sociobiol.* 64, 81 (2009).
- [31] T.B. Ryder, D.B. McDonald, J.G. Blake, P.G. Parker, & B.A. Loiselle, *Proc. R. Soc. B* 275, 1367 (2008).
- [32] <http://people.maths.ox.ac.uk/~porterm/data/facebook100.zip>
- [33] A.L. Traud, P.J. Mucha, & M.A. Porter, *Physica A*, 391, 4165 (2012).
- [34] L. Subelj & M. Bajec, *Physica A* 390, 2968 (2011).
- [35] C.R. Myers, *Phys. Rev. E* 68, 046116 (2003).
- [36] J. Leskovec, J. Kleinberg, & C. Faloutsos. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), (2005).
- [37] T. Opsahl, & P. Panzarasa, *Social Networks* 31, 155 (2009).
- [38] J. Leskovec, J. Kleinberg and C. Faloutsos. ACM Transactions on Knowledge Discovery from Data (ACM TKDD), 1(1) (2007).
- [39] C. Stark, B. J. Breitkreutz, T. Regul, L. Boucher, A. Breitkreutz, and M. Tyers, *Nucleic Acids Research* 34, D535 (2006).
- [40] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabasi, *Nature* 407, 651 (2000).
- [41] J. Leskovec, K. Lang, A. Dasgupta, M. Mahoney. *Internet Mathematics* 6, 29 (2009).

[42] J. Leskovec, L. Adamic and B. Adamic. ACM Transactions on the Web (ACM TWEB), 1(1), 2007.